

***Exposing and Sharing the Hidden Archive:
Advancing Scholarship and Academic Collaboration***

A proposal to the Andrew W. Mellon Foundation

8 October 2007

Center for History and New Media
MS 3G1
George Mason University
Fairfax, VA 22030
Phone: 703-993-4524
E-mail: dcohen@gmu.edu, roy@gmu.edu

Internet Archive
116 Sheridan Avenue
The Presidio of San Francisco
San Francisco, CA 94129
Phone: 415-561-6767
E-mail: linda@archive.org, brewster@archive.org

CENTER for HISTORY  and
NEW MEDIA

<http://chnm.gmu.edu/>



Executive Summary

The Internet Archive and the Zotero Project propose a collaborative project to build a structured archival and access environment for the aggregated materials of thousands of scholars. The project would offer a critical back-end extension to the Zotero research platform to provide permanent archiving and personalized file and project materials management. The platform would create a "Zotero Commons"--publicly shared items from the community of scholars--as well as a more circumscribed 'group' sharing option, for individuals who would like to share documents within limited sets of users.

This project requires extensive changes and preparations by both the Zotero Project and the Internet Archive. The Zotero team will add new options and features to their popular Firefox extension and modify the server they are currently building to function as an intermediary to the storage and other functionality provided by the Internet Archive. In addition to readying substantial storage for the large user base of Zotero, the Internet Archive will also create mechanisms for the conversion of documents donated to the Commons into searchable text (which will be returned to the Zotero Commons user donating the document) that will greatly improve the ability of scholars to scan their personal research corpora.

Zotero will add a prominent "Commons" icon to the top left of the Zotero client interface, just below "My Library." If items dropped onto this icon are in the public domain, they will be added to Zotero Commons, and will be searchable along with all other items made shareable by end-users. The items dragged into Zotero Commons would automatically be stored at the Internet Archive's Data Repository, on dedicated Zotero Commons servers. Internet Archive will in turn deposit back to the Commons an OCR version of their documents to enable full-text searchability. In short, the Commons differs from "My Library" in the Zotero client application in two ways: 1) it moves the selected items into long-term storage at the Internet Archive's Data Repository, and 2) it pulls items into a public, searchable space as opposed to the private, local nature of "My Library."

Items donated to the Commons will be in subdirectories of the Commons named for the donors. In addition to encouraging donations to the commons (since those donating will receive credit for their contributions), this feature will also enable users to identify others who are working with and/or annotating the same content, fostering new collaboration opportunities. The benefits to the scholarly community are thus threefold:

- 1) The availability of permanent, persistent archival, offsite storage for long-term management and use of digital content.
- 2) The ability to share resources publicly for easy access by other scholars.
- 3) The simplified discovery of new, related resources and potential collaboration opportunities.

This partnership will create completely novel and much-needed functionality that will serve a vast number of scholars--not only Zotero users but also the larger academic community. The result of the collaborative project will be a seamless system for archiving, public sharing and collaboration within the scholarly community. The project will build on themes of supporting and enabling cyberscholarship that were the focus of the May 2007 NSF/JISC workshop and the upcoming annual Open Content Alliance Meeting.

The project will take two years to fully implement technologically, disseminate to the scholars who will contribute and use the commons, and to reach full production and use levels, though features and functionality will be implemented in stages and made available as soon as they are ready.

1. Introduction: The Hidden Archive and Digital Scholarship

In the last decade, scholars have been blessed with an exponentially growing collection of digitized and online materials to aid their research. Major projects from universities, libraries, museums, non-profits, and corporations have greatly expanded what is available to the researcher. What once required an expensive and time-consuming trip to a remote library or archive--often not feasible for many scholars--has now entered the realm of possibility. Scholars are increasingly consulting collections on the Web.

This mitigation of the problems associated with research before the massive digitization projects and online collections became available is unfortunately only partial, however, as scholars have begun to realize. Many researchers, and especially those doing advanced work in the arts and humanities, still do not have access to the most critical documents for their research--the primary sources that are at the heart of the best scholarship--and are unlikely to find those documents online in the near future. Most archives, libraries, and museums have not been able to undertake expensive, thorough digitization projects.

Yet in many cases these documents *have* been scanned, copied, or photographed *by someone*. Virtually every professor, graduate student, and author maintains major private caches of Xeroxes, scans, and pictures from their various trips to archives, museums, and libraries. This "hidden archive" likely rivals existing online collections, and is in many cases complementary (rather than overlapping) with what has already been digitized and made available. What if there was a way to expose and share this tremendous hidden archive with scholars from around the globe?

Encouraging such widespread sharing would not need to rely purely on the goodwill of the scholars who retain these private archives. These same scholars currently face a serious problem: they are unable to search their personal collections in the same way they can search online archives. The workflow of the scholar typically involves the creation of a "personal library" of relevant research materials for an article or book project, followed by close reading, note taking and annotation, and writing. What if the same process of sharing also prepared a scholar's personal library for digital research?

First and foremost, this project therefore proposes a compelling quid pro quo that will greatly expand the scholarly commons of digitized materials while greatly enhancing the research capabilities of the average scholar. By bringing together the highly complementary strengths of the premier open online library, the Internet Archive, and the leading open research tool, Zotero, we believe we will be able to address five areas of weakness in the current environment for scholarly research:

a) Exposing and Extending the Hidden Archive

b) Searching the Personal Library

c) Enabling Networked References and Annotations

d) Simplifying Collaborative Sharing

e) Facilitating Scholarly Discovery

2.a Exposing and Extending the Hidden Archive

While institutions have made great strides in placing scholarly content online (often with the help of the Andrew W. Mellon Foundation), individual scholars have rarely, if ever, placed similar materials online. The technical process is daunting for all but the most technically proficient, and there has been little incentive to do so. Each scholar is effectively a small silo of research materials, which at best are shared with their most immediate colleagues or coauthors. Most scholars would have no objection to sharing such sources, except for a minority of cases where a scholar is in the midst of a major project and is worried about others stealing their thunder. But in general, making such materials available would enhance their publications and enable new publications by making it possible for others to go and read or cite the original sources of their scholarship.

The lack of exposure for this distributed, hidden archive is unfortunate given how much individual scholars have in their personal (offline) collections. Virtually every scholar with an advanced degree has a trove of photocopies in a filing cabinet (pre-1990s) or digital scans on a hard drive (more recently) from their theses or dissertations. Biographers have scores of letters, pictures, and ephemera. Architects and architectural historians have photographs of buildings from around the world. Scholars of literature have scanned diaries and manuscripts for insights into the writing process of those they study.

Currently there is no easy way to share these documents with the wider scholarly community. Occasionally researchers post files to scholarly listservs, message boards, or local university forums or storage. Yet these methods do not bring to light the hidden archive and make it broadly accessible. Moreover, documents are often relevant in more than one discipline; posting such files to disciplinary sites (such as H-Net, for historians) assumes that a document will only

be of use in a highly circumscribed realm.

What is needed is a true "commons"--a place where any scholar may access public domain or other shareable documents contributed by the widest array of researchers. In the same way that art historians typically pool their slides into one departmental library that can be used by any instructor or researcher on campus, a place is needed for any scholar who wishes to access, use, and cite the combined documentary evidence from their field and related fields.

No such place exists, and this proposal intends to build one with Zotero as the front end for contributions (and in some cases, access as well) and the Internet Archive as the back end. We propose a simple interface and workflow to enable the creation of this commons.

Adding to the Commons

1. The left column of the Zotero interface, which currently holds "collections" (or folders) of items from a user's "Library" of all research materials, will gain a globe icon labeled "Zotero Commons" (see figure 1).

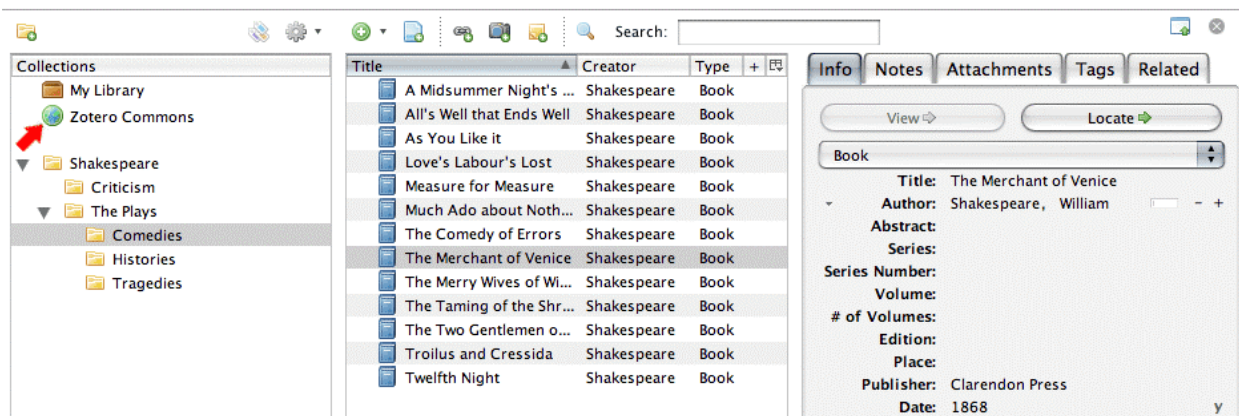


Figure 1: Mock-up of Zotero interface with Commons Icon

2. Documents dragged and dropped onto this icon will trigger:

a) A JavaScript pop-up window with legal caveats emphasizing that the document(s) must be publishable freely on the web and informing the user that the document(s) will be made available as part of the Zotero Commons on the Internet Archive. (See full legal processes below.)

b) A second pop-up window asking the donor to fill in basic metadata, pre-populated with existing Dublin Core fields from Zotero client.

c) A process on the Zotero server whereby the donated document(s) will be associated with the particular user and the document(s) transferred along with this user information and the

document metadata to IA.

d) Items will be stored at IA in ARC format and will have a persistent URL (e.g., www.archive.org/download/SCHOLAR_USERNAME/filename.pdf). Provenance of who has handled the data and versioning will be implemented. If another Zotero user should save the identical document into the Zotero Commons, the item will not be saved/stored a second time; rather the newly stored information will be recorded as metadata and will point to the unique URI of the stored document.

e) A confirmation notice from IA will be sent back to the Zotero user through the Zotero server, including the persistent URL.

Use and Management of the Commons

1. "Owner" of the item: Removal from the Commons will be as simple as clicking on the Zotero Commons icon in the Zotero client to see a listing of items one has placed into the Commons, and then deleting the item as one would in any Zotero collection. (The item will remain in "My Library.") The permanent URL for that item, which may have been cited in the meantime by another scholar, will still exist but a notice will be added stating, "This document has been retracted by the owner."

2. Other Zotero users: We will provide a variety of ways (see below) for other scholars to find items that have been added to the Commons. If a scholar using Zotero wishes to annotate the document or item within the Zotero interface, it will pull from the persistent file location and added to the user's personal collection.

3. Other researchers not using Zotero: IA, as it does for all materials on its servers, will provide a wide array of options for searching and finding items in the Zotero Commons via its web interfaces. The Commons will join other major sections of archive.org such as the web archive, full-text books, films, software, and audio recordings.

4. In addition to regular archiving in the Commons, we will also add any full-text books uploaded by Zotero users to IA's OpenLibrary collection. This will increase the exposure and find ability of full-text items within the scholarly community.

2.b Searching the Personal Library

Most scholars have not yet figured out how to take full advantage of the digitized riches suddenly available on their computers. Indeed, the abundance of digital documents has actually exacerbated the problems of some researchers, who now find themselves overwhelmed by the sheer quantity of available material. Moreover, the major advantage of digital research--the ability to scan large masses of text quickly--is often unavailable to scholars who have done their own scanning or copying of texts.

A critical second part to this alliance of IA and Zotero is to bring robust and seamless Optical Character Recognition (OCR) to the vast majority of scholars who lack the means or do not know how to convert their scans into searchable text. In addition, this process will let others search through such newly digitized texts. After a submission to the Commons, the Internet Archive will subsequently return an OCRed version of each donated document to enable searchability. This text will be incorporated into the donor's local index (on the Zotero client) and thus made searchable in Zotero's powerful quick search and advanced search panes. In short, this process will provide a tremendous incentive for scholars to donate to the Commons, since it will help them with their own research.

IA maintains dedicated servers for OCRing texts. Newly submitted scans will be passed on from the Commons to these servers to be queued for processing. Since there will inevitably be a time delay between the donation of a document to the Commons and the availability of the converted plain text, the Zotero server will serve as an intermediary in this process. IA will return the OCRed text with metadata indicating the user who submitted the document, and the Zotero server will then wait for an established connection with the client software to transfer the text. Indexing will, as always, occur on the client's machine, although the same text will be added into IA's index to enable search across the Commons (see 2.e. below).

Both IA and Zotero understand that even with the incentive of recognition on the Commons and the return of helpful OCRed text many scholars will choose to keep some of their personal library out of the commons. We wish to do something for these scholars as well, so in the second year of the project the two partners will explore the possibility--given adequate remaining storage and other technical considerations--of IA functioning as a key Zotero Storage Provider (ZSP). Zotero already plans in the next version of their software to allow users to specify university storage facilities as the place to backup and provide remote access to their Library. IA could become a primary, or even the default, ZSP in a drop-down menu in the Zotero preferences.

2.c Enabling Networked References and Annotations

One of the pillars of scholarship is the ability for distributed scholars to be sure they are referencing the same text or evidence. As we have already noted, one of the great advantages of the Zotero Commons at IA will be the transport of scholarly materials currently residing on personal hard drives to a public space with stable, rather than local, URIs. These URIs will become critical as scholars begin to use, refer to, and cite items in the Commons.

Yet the partnership this proposal imagines has another benefit: as scholars begin to use not only traditional primary sources that have been digitized but also "born digital" materials on the web (blogs, online essays, documents transcribed into HTML), the possibility arises for Zotero users to leverage the resources of IA to ensure a more reliable form of scholarly communication. Currently the scholarly systems for referencing web pages are in flux, still encountering serious problems with the ephemeral nature of most of the web. For instance, the Chicago Manual of Style recommends to historians and other scholars that they include an "accessed on [this day or during this span of time]" notice at the end of a citation to a web page.

One of the Internet Archive's great strengths is that it has not only archived the web but also given each page a permanent URI that includes a time and date stamp in addition to the URL. Were two historians to reference or annotate a particular web page using the Chicago Manual method, it would remain unclear whether they were really talking about the same thing.

We will address this major problem in several ways. Currently when a scholar using Zotero wishes to save a web page for their research they simply store a local copy. For some, perhaps many, purposes this is fine. But for web documents that a scholar believes will be important to share, cite, or collaboratively annotate (e.g., among a group of coauthors of an article or book) we will provide a second option in the Zotero web save function to grab a permanent copy and URI from IA's web archive. A scholar who shares this item in their library can then be sure that all others who choose to use it will be referring to the exact same document. The Citation Style Language (CSL) code (which formats references in Zotero into a variety of scholarly styles) will be modified to insert these stable URIs and thus make the entire research-to-print process far more rigorous than it stands today.

Moreover, unlike most research software the sophisticated annotation tools built into Zotero--the ability to highlight passages, add virtual Post-It notes, as well as regular notes on the overall document--maintain these annotations separately from the underlying document. This presents the exciting possibility for collaborative scholarly annotation of web pages. Were several scholars to annotate the same archived web page on their individual Zotero installations, they could then aggregate them onto a single document. The Zotero server would act as the go-between for this process, combining individual annotations, while IA would provide the underlying web page on demand.

2.d Simplifying Collaborative Sharing

Groups of scholars also have the need to create more private "commons," e.g., for documents that they would like to share in a limited way. In addition to the fully open Zotero Commons we will establish a mechanism for such restricted sharing. Via the Zotero server, a user will be able to create a special collection with a distinct icon that shows up in the client interface (left column) for every member of the group. The originator of this collection will function as the "administrator," and will create an access control list (with options such as "view only," "add and view," and "add, delete, and view") with the usernames of each group member.

When a member drags and drops a document onto this special collection icon, the Zotero server will receive a message. (If the user is offline, this message will be placed into the Zotero server communications queue and sent when the user goes back online.) The Zotero server will check its access control list for that shared collection to confirm that the particular user has the ability to add items to it, and if so, will send a copy of the submitted document to IA's server for storage. IA will keep these files separate and gated from the exposed and searchable Commons.

After receiving confirmation of a successful transfer, Zotero will send the metadata for that file

to all Zotero users who have access to the particular restricted collection, and an icon for that item will appear in the appropriate folder in their Zotero installations. All requests to view or take action on these shared files will occur through the Zotero server, which will ensure (via a check with the access control list) that the files are not viewed or deleted by a user without the proper privileges.

We believe that having these files reside on the IA server will also encourage the donation of documents at the end of a collaborative project. The administrator of a shared collection will be able to move its contents into the fully open Zotero Commons via a single click in the administrative interface on the Zotero Server.

2.e Facilitating Scholarly Discovery

The multiple libraries of content created by Zotero users and the multi-petabyte digital collections of the Internet Archive are resources that can potentially be of great use to the scholarly community. We believe that neither has experienced the level of exploration and usage we believe is possible through further development and collaboration.

The combined digital collections present opportunities for scholars to find primary research materials, to discover one another's work, to identify materials that are already available in digital form and therefore do not need to be located and scanned, to find other scholars with similar interests and to share their own insights broadly. Yet a repository in its own right, even if open to public access, is not enough to facilitate this kind of collaboration and resource sharing. It is time consuming and labor intensive to slog through gigabytes, terabytes, even petabytes of digitized content to find the relationships and themes that might be of interest to an individual scholar. It is therefore an important aspect of this project that optional discovery features be layered as a service onto rapidly growing collections.

As described earlier, each item contributed to the Zotero Commons, if not already in searchable format, will be converted by the Internet Archive using Optical Character Recognition (OCR) software and contributed back to the Commons via the Zotero server. This process will enable individual scholars to search the entire Commons through a simple text search box. This feature will immediately enable scholars to identify thematically related content, scholars and institutions doing work in similar disciplines, and annotations by other scholars on works of interest. We are not aware of any other open resource today that enables scholars easily to expose and then search for items in a uniform, centralized location. The Commons will accomplish this objective in a completely transparent, automated process.

In the case of websites archived within the Commons, the Internet Archive has the ability to recognize changes in content over time. If another scholar has archived the same site at a different time, the Zotero user will be able to find and compare the sites.

Zotero users are often searching for bibliographic records to support their citations. The proposed project will include development of a function to share bibliographic records. This is a two-step process: 1) the user will be able to search the Internet Archive's OpenLibrary catalog for a

bibliographic record and download it for use in their documents, and 2) the user may be given the option to upload their own bibliographic records to the catalog. This will be a simple API interface to the record upload function being built into the OpenLibrary catalog. The project team will explore these features in the second year of development.

An additional discovery function to be explored by the project team is the possibility of subscribing to RSS feeds (opt-in on both sides) for specific contributors' items placed into the Commons. This would save the individual scholar from having to revisit websites or other resources manually to track the outputs of scholars in related fields or disciplines. This type of time- and effort-saving feature is only one of many that may emerge from real-time use of the Commons as proposed in this project.

3. Development Plan & Timeline

Storage Architecture, API, and workflow

It is expected that only a subset of Zotero users will choose to use the Commons and the archival storage provided through this collaboration. By the end of 2007, there will be approximately 300,000 Zotero users. This proposal envisions an archival storage system of 50TB primary capacity, or approximately 500MB of storage for 100,000 Commons users. Though a single PDF or TIFF file can easily be 50MB in size, we only expect some users to store such large files and believe this is a sufficient amount of storage for the first 2-3 years after service launch at the end of Year 1.

Year 1:

- Install sufficient dedicated hardware at the Internet Archive's Data Repository to support 100,000 users with 500 MB of storage per user. This infrastructure will provide archival storage and serving of materials contributed by Zotero users.
- Add Zotero Commons icon and functionality to the Zotero Commons client application
- Develop API for IA data repository to be interfaced from Zotero Commons icon in Zotero client application
- Adopt server-to-server authentication between Zotero and the Internet Archive using an agreed-upon application
- Monitor usage of the storage feature

Year 2:

- Explore the option of adding IA as a Zotero Storage Provider for archival storage of private My Library collections
- Reassess usage assumptions based on actual usage; consider addition of storage hardware if needed

OCR for Searchability

Year 1:

- Develop and implement automatic OCR transformation of items dragged into Zotero Commons, including return of the OCR text to the Zotero contributor, either as a searchable PDF or as raw text. The OCR text will also be stored on the Zotero server.
- Full text indexing of corpora by applying OCR technology where warranted. This feature may be used to enable search within a Scholars' own Library, and if permission is granted, across multiple libraries. Web content may be indexed separately from the other text such as PDFs, but we anticipate making an integrated interface available to the scholar.

Auto-Discovery (and Recommendation) Functions

Year 1:

- Implement one-click search across IA collections; Search of the Internet Archive's online text catalog with optional display of search results— displayed in web page format or bibliographic records in MARC xml format. The records will be downloadable by the end-user.
- One-click search of the Internet Archive's Archive-It collections and a large portion of the Wayback collection for desired web page content.
- Conversion of web pages not found in the Wayback collections to ARC or WARC file format. The Zotero team and the Internet Archive will work together to allow migration of current and future materials into the standard "warc" format.

Year 2:

- Use deriver to identify other formats of items available in the Internet Archive or Zotero collections
- Automatic generation of reports, to be deposited back with the stored items, that identify other versions of the item that exist in the IA and/or Zotero collections
- Incorporate OCRred text from the Commons into the Zotero recommendations algorithms
- An opt-in function will be provided that generates recommendations of related items in the private My Library collections of Zotero users (to be included in Zotero's release 1.5)

User Interface Enhancements

Year 1:

- Zotero translator searches IA structured metadata for pulling items back into Zotero
- New folders for Commons and group share collections
- Add a link to the Commons from within the Zotero client
- Drop down menu for My Library storage, including IA
- Interface refinement for scholars tailored with input from end-users and tight integration through the Zotero Commons folder, possibly using WebDAV
- User interface enhancement to show all items contributed by an individual (IA implement Zotero Commons server/interface)
- Editing and annotation capability for items stored in Zotero Commons

Year 2:

- Explore option to search by contributor/the whole Commons/discipline or other categories; offer the option subscribe to a feed of materials added by people in fields of interest, designated by Zotero user
- Collaborative annotation window in Zotero

File management

Year 1:

- Storage in ARC/WARC format
- Management via the same processes as all Data Repository Collections
- Creation of persistent URIs for Zotero items to enable stable references.
- Storage of Web material in ARC file format with retrieval (with some display delay) via a version of the Wayback Machine. Other materials would be available as if on a file system stored underneath a web server

Year 2:

- Use metadata to identify multiple references to identical items by different Zotero users (not expected to be a common occurrence)
- Differentiate between book editions, websites that may have changed content, etc, to ensure integrity of Zotero collections (i.e., don't remove duplicate unless item is truly "identical.")
- De-duplicate identical files, if applicable. Implement pointers to items with persistent URIs in metadata for common users
- Versioning, annotation and tracking of some metadata per version of items in the Zotero Commons. Furthermore, these cabinets can be mirrored in other locations (assuming the requestor has the correct file permissions), so this can be a robust system over the long term.

Access Control

Year 1:

- Access control to the Commons will be seamlessly implemented from Zotero client. In the first year, the user will be given two options - fully public through addition to the Commons or fully private by retaining items in My Library.
- Add pluggable authentication to the Zotero server for user logon services and inter-server communications with IA and third parties

Year 2:

- 'Group' sharing option - limited exposure to a set of designated, access controlled users
- Access Control List feature added to Zotero server for private group functionality

Contributions to OpenLibrary and open catalog

Year 1:

- Internet Archive will implement automatic upload of books contributed to the Commons to the OpenLibrary collection.

Year 2:

- Explore addition of a one-click upload function for users to add their own bibliographic records to the Internet Archive's/OCA's OpenLibrary online catalog. This will be an opt-in function.
- The project team will explore the possibility of adding a search and download function to the Zotero client for bibliographic records in the OpenLibrary catalog.

4. Outreach and Dissemination Timeline

Year 1:

- Formally launch the new services resulting from the first year of development, including addition of the Commons feature, automatic generation of OCR'd texts in the Commons, and searchability of the Internet Archive's broader collections to identify items of interest
- Publicize coming features to Zotero users

Year 2:

- Formally launch the second round of services enabled by this collaboration: recommendations, discovery of other scholars with related interests within the Zotero community, etc.
- Perform outreach within scholarly community to raise awareness of Zotero functionality and scholarly discovery and collaboration features
- Do targeted outreach scholarly listservs, newsletters, and annual meetings to find scholars with large troves of documents that could be added to the Commons

5. Copyright Policy

Both the Internet Archive and the Zotero project take copyright extremely seriously. The proposed partnership must balance copyright concerns with the great potential to expose the hidden archive and to foster the kind of collaboration and digital research that would not be possible without this partnership.

Neither the Internet Archive nor Zotero attempts to filter contributions based on copyright issues. Such filtering is not technically feasible, and would likely result in an unnecessary chilling effect on allowable contributions to the Commons. There are no hard-and-fast rules for blocking materials, and no lookup service to check new items against their copyright status. Moreover, a malicious Zotero user could alter the reliable metadata in the "Info" tab of an item to evade any filter.

The Zotero user base, primarily university professors and students, is almost certainly more reliable than the general public and less cavalier with copyrighted materials. But the Zotero project will also have much better identifying information than "rogue" sharing sites do--e.g. IP addresses and usernames that in many cases will be university email addresses. Users intent on sharing copyrighted materials improperly will find IA and Zotero to be environments inhospitable to such activity.

Nevertheless, both projects recognize that some improper materials may be exposed, and have in place strong policies and procedures for dealing with this issue. If a request is made to remove an item from a public collection due to rights status, the Archive has a work process in place for vetting and response. The Internet Archive has been receiving books, audio, and video files from end users for many years to store and host and has had strong and successful policies in place to deal with copyright issues. The processes range from informal dialogue through to consultation with legal experts, and can result in the deletion from public spaces of rights-restricted items.

We also recognize that the average Zotero user may have access to more proprietary or copyrighted materials than the average individual. To thwart both intentional and unintentional sharing of such materials, the Zotero client application will include a submission screen that will be far more emphatic and elaborate than on a commercial sharing site like Scribd or Flickr. In addition to the legal language and checkbox for signing off that an item is shareable, we will include a prominent link at the top to a checklist for assessing whether an item is in the public domain. Below the warning and link, the metadata pane will again emphasize that this isn't your standard (legal or illegal) sharing site (where submissions are often done with little to no metadata); showing fields like the date created will be helpful in making people think twice about improper submissions.

Going the extra mile to warn submitters in this way is a little worrisome since we very well may garner fewer submissions if Zotero users find the submission process onerous. So we will conduct some preliminary focus group studies with a variety of warning page designs to see which one gets across the seriousness of the submission process while also allowing for the easiest process of submission. Moreover, we will strongly emphasize that university storage, rather than the Zotero/IA storage, is the appropriate place for permissible group sharing of gated resources (say, within a university class or department), and will provide simple instructions for doing so. Indeed, this feature will be implemented before the Commons is launched.

Should a Zotero user inadvertently or intentionally deposit a rights-controlled item into the public Commons, the team has ensured that removal is actually easier than uploading that item. The Zotero user need only highlight the offending item and press 'delete'. The item will instantly be removed from the Commons. (The item will remain in the private MyLibrary collection of the Zotero user.)

If a challenge is received by the Archive to an item in the Commons that clearly violates copyright - a filmed television episode, for example - the Internet Archive will immediately remove it from the Commons and notify the up loader. If the status of the challenged item is more ambiguous, Archive staff will work with the individual or organization making the challenge to resolve it satisfactorily. If this dialogue results in taking down the item, the Archive will again notify the up loader.

In some instances where ownership and rights issues are difficult to resolve, the Archive may notify the up loader and, if necessary, engage legal counsel to assist in resolution. This has been rarely, if ever, necessary in the past operation of the Archive's public collections. Nevertheless, it is a backstop option for any particularly difficult issue.

The Archive has posted its formal policy, which reflects the process articulated in the DMCA, on its website, at <http://www.archive.org/about/terms.php>. The Electronic Frontier Foundation and the law firms Cooley, Godward, Kronish LLP and Perkins Coie are the Internet Archive's counsel on these matters. The process is based on the principles and processes of the Digital Millennium Copyright Act (DMCA). (For full text of DMCA see <http://thomas.loc.gov/cgi-bin/query/z?c105:H.R.2281.ENR:>.)

The materials uploaded through Zotero may well increase the load on Internet Archive staff that handles these issues, and this is accounted for in the project budget.

6. Preservation Policies/Technology Migration

The Internet Archive proposes to store, maintain, migrate and preserve fully functional digital master files and metadata for this project.

The Internet Archive's experience in hosting, archiving and providing web access to digitized materials will be leveraged in this project. The Archive therefore has in place numerous processes for risk mitigation and continuation of service. The Internet Archive has evolved preservation and migration policies that are cost effective and scalable to large collections. The preservation strategy includes the storage of multiple copies of files, and at times geographically distributing these. A set of digital master files will be created and archived according to these policies for permanent future access.

The Internet Archive saves two copies of every file created on separate servers in its Data Repository facility in San Francisco. This prevents access interruptions due to failure of an individual machine.

The Internet Archive's technical migration strategy is twofold: hardware and file access formats.

- Hardware: Since its inception in 1996, the Internet Archive has migrated its collections three times – from tape to disk, and twice more to later generation disk systems. Accordingly, technology migration has occurred to and from platforms that have been shown to function at large scale.

- File Access Formats: The Internet Archive tries to maintain accessibility of files by making derivative formats from the original record. The Internet Archive prefers to receive moving image files in MPEG2 format, which was a widely accepted industry standard. Mpeg2 was not an easily accessible format for public use, however, so the files were used to create DIVX versions when that was popular. The original files were later used to create MPEG4 files when that format became popular, and most recently to Flash format. This way the Internet Archive tries to keep files accessible over time.

7. Sustainability

Financial sustainability of the Internet Archive has been demonstrated by eleven years of successful operation as a nonprofit with a broad funding base. This includes fee-for-service web harvesting, book scanning and microfilm digitization, as well as selected federal grants and private foundation grants.

The Internet Archive proposes to store, maintain, migrate and preserve fully functional digital master files and metadata for this project in perpetuity. Since the Zotero Commons files will be stored in the same format as all other collections archived by the Internet Archive, using the same clustered storage environment and management software, these files will automatically be handled with the same preservation and migration policies of all the Internet Archive collections. The value of storing these items at the Archive is that they will, as long as the Internet Archive's collections exist, be maintained with integrity checks and long-term preservation activities.

The Andrew W. Mellon Foundation has generously funded the Zotero project, and although it is in its infancy it is already exploring a variety of means to sustain itself in the coming years. The success of the Zotero community is the most important precondition for a sustainable business model, since the project needs to capture mindshare as well as numerous users in higher Ed who believe the software is indispensable to their scholarly work—and thus want to see further development of the project.

We have envisioned a variety of methods for funding the project long-term given a large user base, including partnerships, services, and support, but have not yet committed to a particular structure. Fortunately, with over 200,000 users and counting the project is making very good progress towards building the community of users and developers, and continues to talk with academics, librarians, and administrators about how Zotero could be sustained for the long run. However, this proposed project has been designed so that even in the unlikely event that Zotero disappears in the distant future the most significant outcome of the collaboration with IA--the Commons of open scholarly material--will remain just as accessible and useful.

8. Budget Summary

The total project funds requested are \$1,166,555 over 2 years as follows:

\$499,055 for Zotero - primarily for development and outreach

\$667,500 for Internet Archive - primarily for development and storage hardware

The allocation of funds over the project period is \$618,153 for year 1 and \$548,402 for year 2.

The largest portion of this project budget will fund software development. The primary tasks include creation of the Commons area and associated features, user interface modifications, synchronization and seamless interaction between the Zotero servers and Internet Archive's data

repository, metadata management and creation of limited access ‘group sharing’ features. Development tasks are described in detail in section **4. Development Plan and Timeline**.

The project budget includes storage hardware for 50TB of primary content at \$100,000. The budget also includes support from Internet Archive staff to resolve copyright issues. A small budget for legal fees has been included to accommodate review of potential copyright issues.

The balance is comprised of small amounts for travel between the two organizations, outreach activities to the scholarly community and other direct expenses. The Zotero team and the Internet Archive are including their entire organizational overhead charges as in-kind contributions.

Detailed budget spreadsheets for both organizations are included as attachments. See Appendices E and F.

Appendix A: About the Internet Archive

“Universal Access to Knowledge.”

The Internet Archive was founded in 1996 to be an “Internet Library,” with the goal of preventing the Internet and other born-digital materials from disappearing into the past. New snapshots of the worldwide web – sites from dozens of countries in over 21 languages – are taken every two months and added to the Archive’s public collection at the rate of 2 Terabytes per month. In the late 1990s the collections were expanded to include books, videos, films, audio recordings and software. The Archive’s collections now consist of over two Petabytes (one petabyte = one million megabytes) of data. The vast majority of collections are available for viewing and download via its public website at www.archive.org.

The Archive is dedicated to open source principles, and has built its storage facilities and its software tools accordingly. The Archive is perhaps best known for the Wayback Machine, a user interface to the historical web archive that enables users to browse websites exactly as they appeared in the past. Many institutions now use the Heritrix web crawler developed and used by the Archive to conduct their own web harvests. The Archive is also known for its large-scale search capabilities – the ability to index and make fully text searchable collections of over half a million items as of late 2006 – using the NutchWAX application. In order to ensure the security and long-term preservation of content, the Archive has partial or full mirror sites around the world, including Amsterdam and Alexandria, Egypt, at the modern Bibliotheca Alexandrina. The storage architecture is simple, low cost and highly scalable.

The Archive acts as a technology partner for large institutions (federal, research, library, international) wishing to create large web collections via web harvesting. It also provides high speed, high quality book scanning services through multiple scanning facilities in North America and Europe, and digitizes microfilm, audio and video collections donated by individuals and partners. Selected partners include the Library of Congress, National Archives and Records Administration (NARA), Biblioteque Nationale de France, the National Library of Australia and numerous University libraries within the US and Canada. The Internet Archive is a founding member of the Open Content Alliance, which has scanned over 150,000 books for free public access, and is a participant in the federally funded Million Books Project.

The Internet Archive, with 11 other international libraries and the Library of Congress, co-founded the International Internet Preservation Consortium (IIPC) in 2003. This international group works to create standards and share best practices to ensure the preservation of ephemeral web content for future generations.

The Internet Archive is funded through service fees from partners as well as grants and corporate partnerships. Selected supporters include the Hewlett Foundation, Sloan Foundation, Mellon Foundation, the State of California, Microsoft and Yahoo. The Internet Archive is certified as a library in the state of California.

Appendix B: About the Center for History and New Media

The Center for History and New Media is well positioned to carry out this project. Since 1994, CHNM (<http://chnm.gmu.edu/>) has used digital media and computer technology to democratize history—to incorporate multiple voices, reach diverse audiences, and encourage popular participation in presenting and preserving the past. CHNM combines cutting-edge digital media with the latest and best historical scholarship to promote an inclusive understanding of the past as well as broad historical literacy. CHNM’s work has been recognized with major awards and grants from the National Endowment for the Humanities (NEH), the Department of Education, the Library of Congress, the Institute of Museum and Library Services (IMLS), the National Historic Records and Publication Commission, and the Sloan, Mellon, Hewlett, Rockefeller, Gould, Delmas, and Kellogg foundations.

CHNM maintains more than four dozen online history projects directed at diverse audiences, making them available at no cost through its website. In 2006, CHNM’s websites had 300 million hits and 10 million visitors—making it one of the busiest non-commercial history education sites on the Web. CHNM’s staff of more than forty people (including eleven with doctoral degrees) includes those skilled in all aspects of digital history—database architecture, software engineering, web development and design, preservation, and archiving. CHNM has an annual budget of more than \$1.5 million and a \$2 million endowment (achieved with the assistance of a \$500,000 Challenge Grant from NEH), which guarantees the long-term stability of CHNM.

CHNM has been a leader in using digital media to improve the understanding of history by students and the general public; in helping libraries and museums explore new and more effective ways to use digital materials and methods; and in building innovative, easy-to-use digital tools for accessing historical content. CHNM maintains several very popular and award-winning web-database projects, including educational projects such as History Matters, Liberty, Equality, Fraternity: Exploring the French Revolution, World History Matters, and Historical Thinking Matters as well as collections-based public history projects such as the September 11 Digital Archive, the Hurricane Digital Memory Bank, and Gulag: Many Days, Many Lives. CHNM is also the lead partner in four major Teaching Traditional American History grants from the Department of Education and a partner in two others. CHNM has won (or shared) the American Historical Association’s James Harvey Robinson Prize three times for its “outstanding contribution to the teaching and learning of history.”

CHNM also has extensive experience working with museums and libraries on digital projects. For example, CHNM worked closely with the Smithsonian Institution’s National Museum of American History to provide an interactive “tell your story” component for NMAH’s award-winning “September 11: Bearing Witness to History” exhibit. We have also worked closely with the Library of Congress, which accessioned the September 11 Digital Archive as its first major digital acquisition. In partnership with the University of Maryland, we have received a grant from the Library’s National Digital Information Infrastructure and Preservation Program (NDIIPP) to preserve the digital history of dot-com boom.

Finally, CHNM has emerged as a leading developer of free tools and software for scholars, librarians museum professionals, educators, and students. Through the Echo: Exploring and Collection History Online—Science, Technology, and Industry project funded by the Sloan Foundation as well grants from NEH, IMLS, and the Mellon Foundation, CHNM's freely available tools have attracted hundreds of thousands of users. CHNM is also currently developing digital tools for teachers and students under a grant from NEH's EDSITEment program.

Appendix C: About the Zotero Project

Zotero (<http://www.zotero.org>) is an innovative, open-source reference management, note-taking, and research tool already funded generously by the Andrew W. Mellon Foundation. Zotero helps users gather, organize, and analyze sources (citations, full texts, web pages, images, and other objects), and lets users share the results of their research in a variety of ways. An extension to the popular open-source web browser Firefox, Zotero includes the best parts of older reference manager software (like EndNote)—the ability to store author, title, and publication fields and to export that information as formatted references—and the best parts of modern software and web applications (like iTunes and del.icio.us), such as the ability to interact, tag, and search in advanced ways. Zotero integrates tightly with online resources; it can sense when users are viewing a book, article, or other object on the web, and—on many major research and library sites—find and automatically save the full reference information for the item in the correct fields. Since it lives in the web browser, it can effortlessly transmit information to, and receive information from, other web services and applications; since it runs on one's personal computer, it can also communicate with software running there (such as Microsoft Word). And it can be used offline as well (e.g., on a plane, in an archive without WiFi).

Zotero already provides advanced functionality for gathering, organizing, and scanning one's research, as well as significant import/export capabilities (including integration with Word and an API for communication with any program or service on the web). Zotero users will soon gain the ability to share and collaborate on their collections with other users through an exchange server, and receive recommendations and feeds of new resources that might be of interest to them.

In under a year since its initial beta release in October 2006, over 200,000 researchers in 15 languages have adopted it, with especially high usage at colleges and universities. It was named among PC Magazine's "Best Free Software" applications for 2007, and received the American Political Science Association's "Best Software" award at the Association's September 2007 meeting. The Mozilla Corporation, recognizing the importance of Zotero for scholarly research, released a special edition of their Firefox browser (Firefox Campus Edition) with Zotero preinstalled in August 2007.

Appendix D: Overview of the Internet Archive's OpenLibrary Project

The Internet Archive has begun a project to build an open source, open access, structured resource that provides information about and access to library resources around the world. The purpose of this project is to develop a free, publicly accessible website that has a web page for every published book, with as much information as possible about that book. It will be a gateway for citizens to the now daunting bibliographic and scanned book repositories that exist and are being created. The need for such a resource has emerged from the Internet Archive's work with the libraries of the Open Content Alliance and other institutions. Other online book initiatives underway do not offer the comprehensive, unrestricted, open access that this resource will.

The OpenLibrary site will serve as a finding and access tool for digitized books and information about all published books. It can serve as an open union catalog for libraries that cannot afford to include their entire collections in other, closed ILS or OPAC systems. The website will have tools to enable libraries, historical societies and other institutions and individuals to upload, organize and access information about books, whether in physical or digitized form. It will enable libraries to more easily share records, to organize un-catalogued collections, and to facilitate digitization of high-demand collections. Bibliographic records can be up- and down-loaded without restrictions on reuse. The website is expected to be of particular value for access to books that may be held in remote, rural, or otherwise difficult to reach collections.

The site will be built on Wikipedia architecture with one page per published book. The initial corpus of pages and books will be based on the Library of Congress catalog and ONIX feeds from publishers and digitized books provided by the Open Content Alliance, the Internet Archive, and other partner libraries with digitized volumes. Each page will have pointers (links) to library websites that have physical or digitized copies of the book and the text of the scanned book, if available. Pages may also have links to publisher and e-tailer websites, enabling patrons to purchase books if they wish. A discussion, analysis and interpretation community is expected to grow around each page, using the blog spaces enabled by the site. A development version of the site architecture is in operation now, and has already begun to attract significant contributions from developers and book enthusiasts.

The State of California under the LSTA program and the Internet Archive are currently funding this project.